Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 16 (2014) 281 – 287

Procedia
Economics and Finance

www.elsevier.com/locate/procedia

21st International Economic Conference 2014, IECS 2014, 16-17 May 2014, Sibiu, Romania

Multidimensional Data Analysis - Representation, Security and Management

Raluca-Mariana Ștefan[a,*], Mariuța Șerban[b], Costin Rudăreanu[a]

[a]*Academy of Economic Studies, Bucharest, Romania*
[b]*Pitesti University, Pitesti, Romania*

Abstract

Providing correct and specific information for economic management can be achieved through an object identification system and can be guaranteed in order to get correct management decisions through applying certain security procedures. This paper describes the principal components analysis as a method of data representation and multidimensional data reduction and also the results obtained when applying this technique and the K-means clustering algorithm on a set of economic data. At the end, there is described the projection of the cluster analysis through the means of electronic security and economic management.

Selection and/or peer-review under responsibility of Scientific Committe of IECS 2014
Keywords: multidimensional data, data representation, principal components analysis, data security, data management

1. Introduction

The systems that define a series of activities necessary within an object recognition process, the information sets, the procedures, the algorithms and the techniques used for this purpose are known as object recognition system. Depending on how the assignment of an object to a certain group is done, there are two types of object recognition systems: unsupervised learning systems and learning systems. The targeted objectives can be accomplished by using one of the identification systems and its methods that corresponds to the data and information type.

* Corresponding author.
E-mail address: rstefan2012@yahoo.com (R. M. Ștefan), mariuta_serban@yahoo.com (M. Șerban), costin.rudareanu@yahoo.com (C. Rudăceanu)

Selection and/or peer-review under responsibility of Scientific Committe of IECS 2014
doi:10.1016/S2212-5671(14)00802-8

Unsupervised learning or cluster analysis involves only data which are unlabeled (Duda et al., 2001), which are then grouped according to their natural tendency. An unsupervised system of automatic learning is used to summarize information for extremely large volumes of recorded digital data that also has a high level of heterogeneity and complexity.

Consequently, unsupervised automatic learning procedures are useful and effective. The main feature of unsupervised learning systems is that the belonging of the analyzed subjects to a specific group of objects it is not known, which can indicate that the number of groups for the considered objects is also unknown until the end of the data clustering. In order to perform data clustering their representation plays a very important role and the principal components analysis is one of the most used techniques for representing multidimensional data by dimensionality reduction.

In the literature, PCA is considered to be the continuous solution to k-means clustering algorithm and the subspace spanned by the PCs is identical to the subspace defined by the cluster centroids (Ding, He, 2011). Pattern identification in data and highlighting data similarities and differences are just two of the PCA technique advantages. In this paper, it was considered a data set and PCA and K-means algorithms were applied in order to compare the results.

In the literature, PCA is considered to be the continuous solution to k-means clustering algorithm and the subspace spanned by the PCs is identical to the subspace defined by the cluster centroids (Ding, He, 2011). Pattern identification in data and highlighting data similarities and differences are just two of the PCA technique advantages. In this paper, it was considered a data set and PCA and K-means algorithms were applied in order to compare the results.

2. Sample of Multidimensional Economic Data Set

The data set consists in 10476 instances regarding daily energy consumption and it was taken from UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, the set being available at https://archive.ics.uci.edu/ml/datasets.html.

Attribute information for the data set: global active power (household global minute-averaged active power in kilowatt); global reactive power; household global minute-averaged reactive power in kilowatt; voltage (minute-averaged voltage in volt); global intensity (household global minute-averaged current intensity in ampere); sub metering 1 (energy sub-metering No. 1 in watt-hour of active energy and it corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave); sub metering 2 (energy sub-metering No. 2 in watt-hour of active energy and it corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light); sub metering 3 (energy sub-metering No. 3 in watt-hour of active energy and it corresponds to an electric water-heater and an air-conditioner).

3. Multidimensional Data Representation by Principal Component Analysis

The principal component analysis and the Karhunen – Loève transformation have a strongly connection and represent classic statistical techniques used for data analysis, characteristics extraction and reduction of data dimension.

In order to create a proper image of the changing nature of an economic process or phenomenon it is commonly used the concept of variable. Variable is an abstraction of the possible set of values that a data characteristic can register (Ruxanda, 2001).

Being given a cluster of multidimensional data, the purpose of their analysis is to find a smaller cluster of variables that can offer a very precise representation of the initial cluster. For the principal components analysis, redundancy is measured by the correlations between the elements.

Among the variables of a data set there are connections that are linear or they can be assimilated to a linear connection so that principal components analysis (PCA) can be applied. By this technique initial data are transformed to new data that are more suitable to be analyzed. Applying PCA to a set of data having two or more variables, this procedure generates a new set of data that has the initial number of variables called principal components.

Each of these principal components consists in a linear transformation of the initial data cluster and their coefficients are calculated so as to contain the variances in the appropriate descending order. There are at least two principal components and the principal components are not linearly correlated.

Consider vector X with n elements.

Given vector XM composed of elements of a random vector X, respectively X(1) , ... X(M). Principal components analysis does not require generating hypothesis on the probability density of the vectors if the first and second order statistics are known or can be estimated.

The elements of the vector X are correlated with each other and there is the possibility of dimension reduction of X. If the elements of vector X are independent then the principal components analysis cannot be applied.

Data processing before applying the technique of principal components analysis involves data centering and data standardization. Data centering is performed by replacing them with the difference between the original items and their average.

In most cases, the data used in PCA needs to be standardized by substituting the sample mean from each observation and dividing it by the sample standard deviation.

PCA consists of two steps as it follows:

I. The covariance matrix of the considered data is calculated;

II. The eigenvectors are computed and they correspond to the principal components of variation in the data.

Because the original data can be recovered in its main components PCA exact form, it is a very useful tool for reducing data dimensionality, for visualizing data etc. The principal component coefficients and the variances that correspond through the eigenvectors functions of the covariance matrix are then calculated.

After PCA was applied for dimension reduction there were obtained principal components which contain as much information as possible from the initial data set and these are displayed in Figure 1.

Depending on the results, the sequencing of the principle components in the descending order of the obtained values is carried out as follows: the highest of the values will correspond to the first principal components; the next value will correspond to the second principal component and so on.

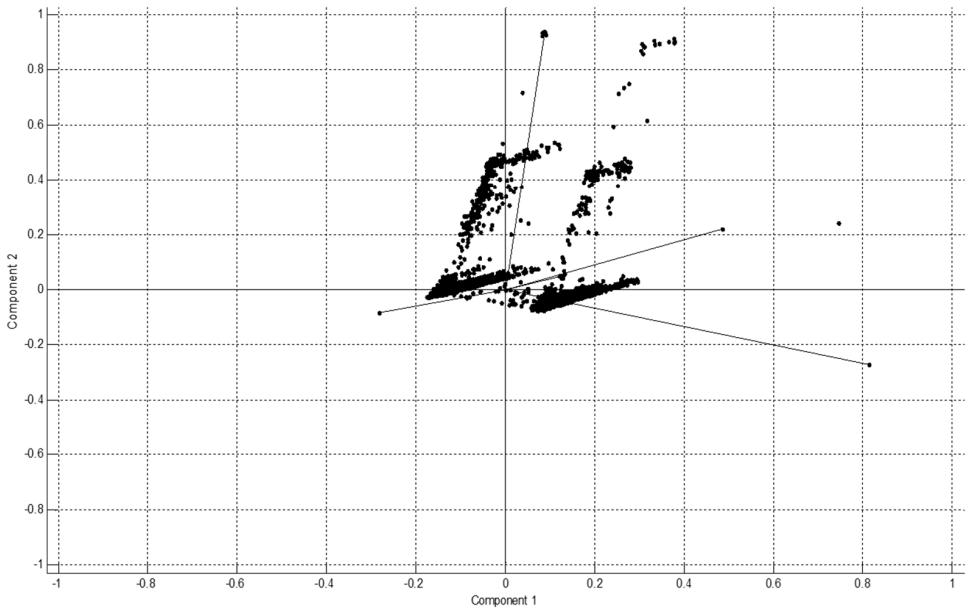

Figure 1: Principal components

Source: Authors' results from MATLAB

PCA main goal is to extract from a multidimensional data set the smallest number of components, principal components, which can hold original information as much as possible.

Applying partitional methods over a set of data that was reduced implies less time for processing due to the lower number of data attributes and data are clustered with a higher level of accuracy. One of the most used partitional algorithm is K-means.

4. K-means Clustering

Because economic data is defined by a very low level of homogeneity, performance prediction based on this type of data can be improved if it is based on decisions made by experts, substantiated with the help of cluster analysis techniques.

Clustering extremely large data sets using cluster analysis represents an exploratory technique for obtaining information relevant to the decision-making factors which can then use this knowledge to decide accordingly, short term or long term, with the purpose of improving economic performance.

One of the advantages of using a partitional algorithm is that the solution offered is one of unilevel type, which determines its application on a large set of data. Another advantage resides from the fact that the number of clusters obtained from the data set may be subsequently changed, so that in the end the optimal solution is found. These are the main features of a partitional clustering algorithm. Being a partitional clustering method, K-means algorithm allows for the number of clusters to be proposed from the beginning and the number k can be changed for optimal cluster dividing. In order to find the most performing and efficient number of clusters, the algorithm is applied over the data set for a few times.

K-means method of clustering data represents a partitioning procedure which supposes that the data are objects and the distances between them and their locations group them. K-means separates the objects into k mutually exclusive clusters, so that the distances between objects within a cluster are as small as possible and the distances between objects assigned to different cluster are as big as possible, having for each cluster its centroid that represents the cluster center point.

Initially, for the data set that was considered for application, the K-means algorithm has been applied for k=2 and the results can be viewed further on. The two projected clusters contain the clustered objects and are distinguished by a triangular or square shape in Figure 2.

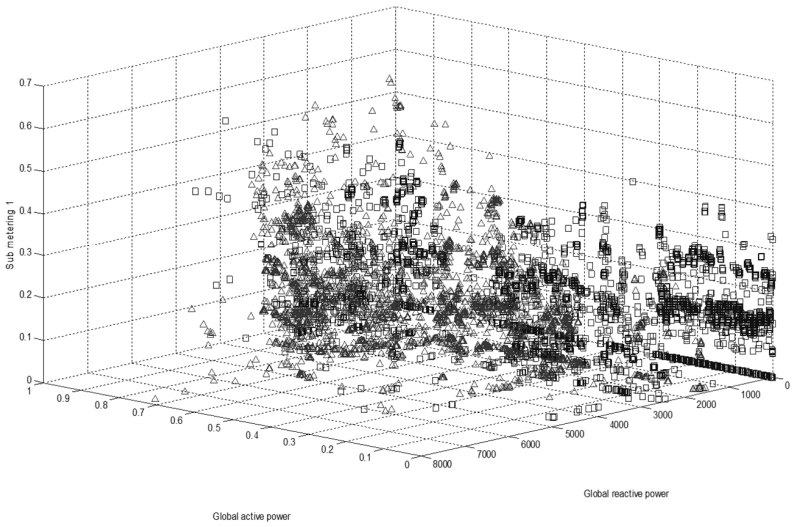

Figure 2: Clustering multidimensional data by K-means algorithm
Source: Authors' results from MATLAB

The silhouette graph has been determined for result observation in the separation of the two resulted clusters. The values can be situated in the range [-1,1]. This graphical method performs both the interpretation and the validation of the clusters obtained by providing a representation, relating to each item's affiliation to the cluster. The existence of a negative value in a silhouette graph for one of the cluster's objects would indicate a high probability of the fact that the object does not belong to the cluster to which it was assigned.

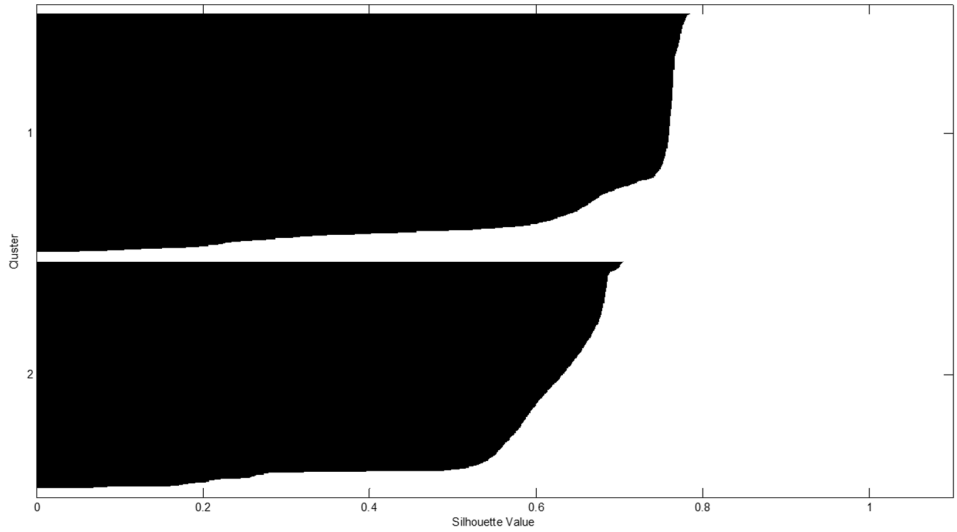

Figure 3: Silhouette plot (k=2)
Source: Authors' results from MATLAB

From Figure 3 it can be noticed that the separation of the two clusters is very good, indicating that the K-means algorithm has made a correct clustering of the considered data and the values of the first cluster are close to 0.8. There are also points that have lower silhouette values, which indicate that they are located close to the points belonging to the other cluster. Also, the absence of negative values corresponding to clustered objects indicates a correct assignment of the objects designated to the respective clusters.

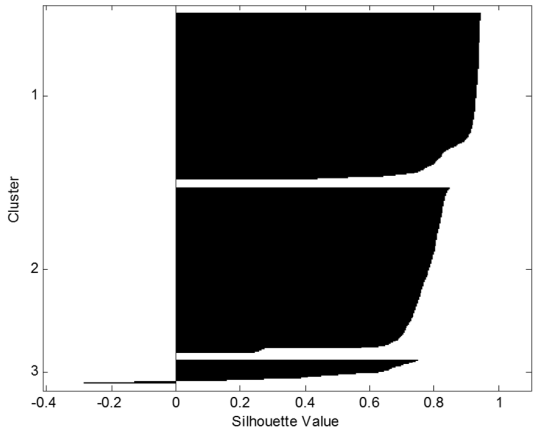

Figure 4: Silhouette plot (k=3)
Source: Authors' results from MATLAB

From Figure 4 there can be noticed some negative values which indicate that the objects corresponding to these values were assigned to a cluster other than the one that they actually belong to. Also, the majority of the values are below 0.8 which indicates that the assignment of three clusters of considered data is not as good as the one carried out in two clusters. Therefore, the K-means partitional algorithm offers a considerable advantage which allows changing the number of clusters until data clustering is performed in an optimal manner.

5. Multidimensional Economic Data Security and Management

A multidimensional data security and management imply to protect information and information resources of an economic entity in compliance with the policy requirements, wherever this information is stored, processed and/or communicated. Technology offers many solutions but it is necessary for an efficient management to utilize other forms of protection in order to diminish risks, vulnerabilities and attacks to information.

By using a multidimensional and multilevel security system, databases are secured so that any information extracted from them is accurate and produces knowledge that can be used in management decisions. Permissions of users or groups of users are used in order to manage access to data and objects. The groups of users are called roles. In Microsoft SQL Server, permissions for data or objects are specified by roles.

The users in a role that has permissions for a particular object can use that object and have equal permissions to the objects. Each object has a permissions collection with the permissions granted on that object, different sets of permissions can be granted on an object and for each of the permissions from the permissions collection of the object exists a single role assigned to it (http://technet.microsoft.com).

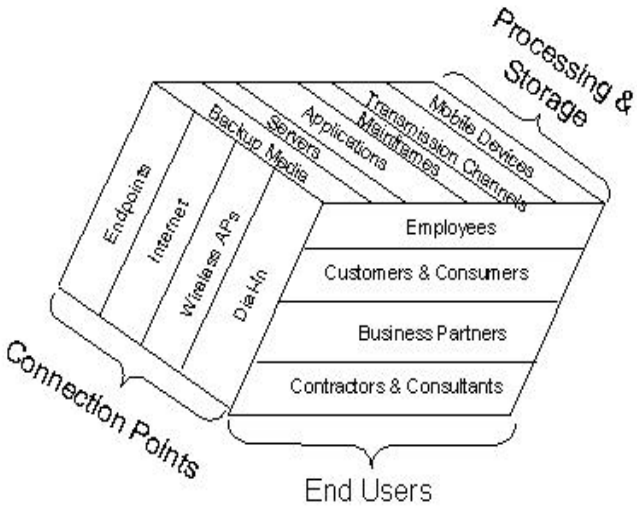

Figure 5: Information Components of an Economic Entity
Source: Herold, R. (2006) The Definitive Guide to Security Inside the Perimeter, RealTimeNEXUS

The Digital Library for IT Professionals; available at http://nexus.realtimepublishers.com/dgsip.php.

6. Conclusions

Multidimensional data analysis contains methods and techniques that are considered the most appropriate tools used to identify similar or dissimilar structures that can be applied in various fields of the economic segment.

The delivered results indicate that reducing the data dimension produces a significant reduction in the processing time and this way it is also possible to increase the performance of the applied techniques.

The data analysis process is an informational type of process that has as input data, the primary data and as output data or results, complex data, summarizers.

By applying cluster analysis procedures to a set of objects, each object will be assigned to a unique cluster and the formed clusters will have a generalizing significance based on which there can be drawn conclusions for the knowledge process. Therefore, the security of the data used in the cluster analysis is very important for the management of each economic entity.

The significance and the relevance of the results are directly related to the accuracy of the procedures applied to the data and choosing one method or another depends on the characterization and the representation of the data.

The accuracy of the decisions is also given by the concise knowledge provided after performing the analysis conducted through cluster analysis procedures and the increased security of the used data. Redundancy of data base collected information represents a major challenge for data clustering and for database security, thus requiring the creation of models and methodologies which have very good results in terms of efficiency and correctness.

References

Chae S., Warde W. (2006) Effect of using principal coordinates and principal components on retrieval of clusters, Computational Statistics and Data Analysis, Volume 50, Elsevier B.V., pp. 1407-1417.

Ding, C.; He, X. (2004) K-means Clustering via Principal Component Analysis. Available at http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf.

Herold, R. (2006) The Definitive Guide to Security Inside the Perimeter, RealTimeNEXUS The Digital Library for IT Professionals; available at http://nexus.realtimepublishers.com/dgsip.php.

http://technet.microsoft.com.

https://archive.ics.uci.edu/ml/datasets.html.

Jain, A.K., Dubes, R.C. (1988) Algorithms for Clustering Data, Prentice Hall Inc., 320 pp.

Napoleon, D.; Pavalakodi, S. (2011) A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set, International Journal of Computer Applications (0975 – 8887), Volume 13– No.7, January 2011.

Ruxanda G. (2001) Analiza datelor, Ed. ASE, Bucureşti.

Ruxanda, G. (2009) Multidimensional data analysis – Doctoral School Course Notes, 133 pp.

Sembiring, R.W.; Zain, J.M.; Embong, A. (2011) Dimension Reduction of Health Data Clustering, International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(3): 1041-1050, The Society of Digital Information and Wireless Communications, ISSN: 2220-9085.

Tibshirani R., Walther G. (2005) Cluster Validation by Prediction Strength, American Statistical Association, Institute of Mathematical Statistics and Interface Foundation of North America, Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 511-528.

www.mathworks.com.